# 基于 SimHash 和混合相似度的多模式匹配方法 *

曹卫东，胡 炜，王家亮，王 静

(中国民航大学 计算机科学与技术学院，天津 300300)

**摘 要：**为了解决多源异构民航旅客服务数据集成过程中存在多模式匹配的效率不高、精确性不足、完整模式信息获取难度较大等问题，提出了一种基于 SimHash 和混合相似度的多模式匹配方法。该方法首先基于 PMI 计算特征单元权重，并通过 SimHash 算法构造属性列的签名来表示属性特征，以降低特征维度，进而引入 K-means++算法对属性聚类并生成候选匹配集。最后基于属性的混合相似度构建属性映射图，以直观的方式展示属性间的匹配关系，同时提高多模式匹配效率。实验结果表明该方法具有可行性，为高效的解决多源异构民航旅客服务数据集成中的模式冲突问题提供新的解决方案。

**关键词：**多模式匹配；签名；点互信息；混合相似度；属性映射图

**中图分类号：**TP391 doi: 10.19734/j.issn.1001-3695.2018.06.0462

## Multiple schema matching method based on simhash and mixed similarity

Cao Weidong, Hu Wei, Wang Jialiang, Wang Jing

(*College of Computer Science &Technology*, *Civil Aviation on University of China*, *Tianjin* 300300, *China*)

**Abstract:** In order to solve the problems of multiple schema matching in the process of integrating multi-source heterogeneous civil aviation passenger service data, such as low efficiency, low accuracy and the complexity of obtaining complete schema information, this paper proposed the multiple schema matching method based on SimHash and mixed similarity. Firstly, the method calculated the weight of feature units based on PMI, and generated the signature of columns by SimHash to represent attribute features to reduce feature dimension. Further, it employed K-means++ to generate candidate matching sets by clustering the columns. Finally, it constructed the mapping graph of attributes based on attributes' mixed similarity, and displayed the matching relationship between attributes intuitively. Meanwhile, it improved efficiency of multiple schema matching. The experimental results verify the feasibility of the proposed method. The method provides a new solution for efficiently resolving the schema conflict in the process of integrating multi-source heterogeneous civil aviation passenger service data.

**Key words:** multiple schema matching; signature; PMI; mixed similarity; attribute mapping graph



## 0 引言

国内民航旅客服务信息系统中存有大量与收益相关的数据，如 PNR(Passenger name record,旅客姓名记录)、ET(Electronic Ticket,电子客票)、CKI(Check-In,离港信息)等。这些数据由不同应用产生并分散在各自的系统中,存在模式异构问题。如果对民航收益漏洞产生的原因进行综合分析,就需要对多源异构数据进行集成。而模式匹配技术是数据集成的关键技术,它可以发现属性之间的语义映射关系,消除数据模式的异构冲突。

目前,因民航旅客服务领域数据安全性要求较高,数据的多级安全访问权限造成数据的详细模式信息获取较难,经分析传统的模式匹配方法应用在此领域存在如下问题:

a)基于模式信息的匹配方法对模式信息的完整性要求较高，当模式信息不全时,仅通过计算属性之间的文本相似度来区分属性,无法获得理想的匹配结果。如表 1 中的 psg_type 属性与表 2 中的 opt_type 属性之间的文本相似度很高,但它们指代意义却全然不同。

b)基于数据实例的匹配方法无法解决因数据分布特征相似带来的误配问题。如表 1 中的 orgn_city 代表出发城市,表 2 中的 destination 表示到达城市，它们的数据实例分布相似,但若据此判断它们表示相同的语义就会导致错误。

c)传统方法解决的是二元匹配问题，当 $n$ 个数据模式进行

---

**收稿日期：**2018-06-20；**修回日期：**2018-08-21 **基金项目：**民航局科技创新引导资金重大专项资助项目（MHRD20150107，MHRD20160109）；中央高校基本业务费资助项目（3122014C017）

**作者简介：**曹卫东（1964-），女，天津人，副教授，博士，主要研究方向为数据库与数据挖掘、民航信息系统软件可靠性（wdcaomiss@163.com）；胡炜（1993-），男，河南信阳人，硕士研究生，主要研究方向为航空运输大数据、数据集成；王家亮（1983-），男，辽宁辽阳人，讲师，博士，主要研究方向为嵌入式系统，民航信息系统；王静（1980-），女，山西晋中人，讲师，博士，主要研究方向为民航信息系统、大数据.

匹配时,传统方法每次匹配两个模式,需要迭代 $n(n-1)/2$ 次,因此匹配效率不高。

表 1　民航旅客订座数据表(PNR)

Table1 Civil aviation passenger booking records data sheet

| tk_no | orgn_city | psg_type | … |
|---|---|---|---|
| 9．99979E+12 | TAO | CIP | … |
| 9．99242E+12 | TAO | VIP | … |
| 9．99852E+12 | CGO | CIP | … |
| 9．99242E+12 | SHE | CIP | … |

表 2　电子客票数据表(ET)

Table2 Electronic ticket data sheet

| opt_type | destination | TICKET_NO | … |
|---|---|---|---|
| R | CGO | 9．99242E+12 | … |
| R | BZV | 9．99242E+12 | … |
| V | SSG | 9．99852E+12 | … |
| R | TAO | 9．99442E+12 | … |

针对上述问题,本文提出一种基于 SimHash 和混合相似度的多模式匹配方法来探究属性之间的匹配关系,将模式级的匹配问题转换成属性级的匹配问题,在简化多模式匹配过程的同时提高了匹配质量。

## 1　相关工作

模式匹配方法经多年发展,已经取得了较好的成果[1],根据辅助匹配信息的不同主要分为以下四种:

a)基于模式信息的匹配,如 COMA[2]和 SEMINT[3]通过结合多种模式信息,取得了较好的匹配效果,但当多个模式匹配时,此类方法的时间复杂度较高。Ding 等人[4]提出通过 TF-IDF 方法构造属性的特征向量对属性聚类分析降低了多模式匹配的时间复杂度问题,但存在"同名异义"和"同义异名"的属性,导致其匹配效果不佳。

b)基于结构信息的匹配。早期的基于结构信息的模式匹配方法将源模式和目标模式通过树或图的结构表示出来,再通过计算对应节点之间的相似度来挑选最佳匹配[5-6],为了提高模式匹配的准确度,后来杜小坤等人[7]又提出了 IU_Based 方法。

c)基于数据实例的匹配:早期基于数据实例的匹配方法。通过获得数据实例的重复度来挖掘属性之间的匹配关系[8, 9],该类方法挖掘出的数据实例分布特征并不完整。Ahmadi 等人[10]提出用 q-gram 方法结合互信息理论构造属性列的特征向量,但当数据实例相似时会导致误配情况(例如数值型的属性列之间无法区分);Mehdi 等人[11]提出通过正则表达式来匹配数据实例的方法就解决此问题,但是该方法最后仅借助谷歌相似度来区分误配属性,准确度受限;Gu 等人[12]提出将实例匹配和模式匹配交互执行,从而提高模式匹配的准确度也是一种可行的方法,但匹配过程趋于复杂。

d)基于其他信息辅助匹配的方法。如运用本体知识构造模式本体与全局本体进行匹配的方法[13],为解决多模式匹配问题提供了新的思路,但在实际匹配过程中需要的多种模式信息不易获得。

在数据具有访问权限,无法获得详细模式信息的背景下,以上四类方法难以获得理想的匹配效果。因此本文提出使用属性的数据实例和属性名作为辅助匹配信息的方法,该方法在数据安全要求较高的民航领域同样适用。

## 2　基于 SimHash 的聚类分析

### 2.1　基于点互信息的 SimHash 算法

在多模式背景下,属性的数据实例又称为属性列。由于属性列与属性是一一对应的关系,因此可以用属性列的特征来表示属性特征。传统方法用互信息理论来计算属性之间的相似度时,由于提取的特征单元较多,导致属性列的特征向量维度较高,不利于后续的计算。本文使用基于点互信息(PMI)的 SimHash 算法来生成固定位数的签名作为属性列的特征,有效的降低了特征向量的维度。并给出如下相关定义。

定义 1　特征单元。指从数据实例中提取的具有实际含义可以用来表示数据实例特征的数值或者字符串。

由于数据实例复杂多变,因此将数据实例分为字符串型、时间型和数值型三类,以便提取特征单元。跟据定义 1 对字符串类型的数据用 q-gram 提取特征单元。时间型数据按照年、月、日、时、分、秒等进行单位分割处理。数值类型具有稀疏性,可以采用等距划分法提取特征单元。属性列 $a$ 提取特征单元后,以键值对的形式表示为 $a=\{<u_1,ta(u_1)>,<u_2,ta(u_2)>,\cdots,<u_i,ta(u_i)>\}$,其中,$u_i$ 是 $a$ 的特征单元,$ta(u_i)$ 是 $u_i$ 在属性列 $a$ 中出现的频次。$n$ 个属性列的所有特征单元的交集为 $U=\{u_1,u_2,u_3,\cdots,u_m\}$,代表属性列的特征集合。

定义 2　点间互信息。指衡量任一属性列 $a_k$ 与任一特征单元 $u_y$ 之间所蕴涵信息量差异的一种量度,用 $\mathrm{pmi}(a_k,u_y)$表示。

$$\mathrm{pmi}(a_k,u_y)=lb\frac{ta_k(u_y)/T}{\left(\sum_{i=1}^{n}ta_i(u_y)/T\right)\times\left(\sum_{y=1}^{m}ta_k(u_y)/T\right)} \tag{1}$$

其中: $ta_k(u_y)$ 表示特征单元 $u_y$ 在属性列 $a_k$ 中出现的频次, $\sum_{i=1}^{n}ta_i(u_y)$ 表示特征单元 $u_y$ 在所有属性列中出现的频次和, $\sum_{y=1}^{m}ta_k(u_y)$ 表示特征集合 $U$ 中的所有特征单元在属性列 $a_k$ 中出现的频次和,$T$ 表示所有特征单元在所有属性列中出现的频次和。

SimHash 算法[14]是一种计算海量文本相似度的高效算法,其原理是将高维文本特征转换为固定位数的签名,通过比较签名来获取相似关系。由定义 2 可知,属性列与其包含的特征单元之间的 PMI 值越大,则该特征单元与当前属性列的相关相关性越大。若两个属性列相同的特征单元越多,则这两个属性列匹配的可能性越大。因此本文使用特征单元与属性列的 PMI 值作为权重,提出基于 PMI-SimHash 的属性列签名生成算法。

算法 1　生成属性列签名

输入:$n$ 个属性列集合 $A=\{a_1,a_2,\ldots,a_n\}$。

输出:所有属性列的签名集合 $P$。

```
P=Ø
for a∈A
    a={u1,u2,…,uy}
    S=0
    for u∈a
        s=hash(u)
        if si==0
            si= pmi(a,u)
        else
            si= -pmi(a,u)
        end if
        S=+s
    end for
    if Si>0
        Si=1
    else
        Si=0
    end if
    P=P.add(S)
end for
return P
```

具体步骤为:遍历属性列集合 $A$,提取属性列 $a$ 的特征单元集合 $a=\{u_1,u_2,\ldots,u_y\}$;用相同的 hash 函数生成特征单元 $u$ 的 $f$ 位签名 $s$,并根据式(1)计算 pmi($a,u$)。若 $s$ 中第 $i$ 位为 1,则更新 $s$ 的第 $i$ 位为 pmi($a,u$); 否则,更新为-pmi($a,u$), 并对属性列 $a$ 中的所有特征单元 $u$ 的签名 $s$ 进行按位求和得 $S$,判断 $S_i$ 并更新其值; 最后将属性列的签名 $S$ 加入到集合 $P$ 中,最后返回所有属性列的签名集合 $P$,算法结束。

### 2.2 聚类分析

由于属性与属性列是一一对应的关系,因此用属性列的签名来表示属性特征并进行聚类分析,即可得出属性的聚类关系。k-means++算法[15]作为一种基于划分的聚类算法其优点在于收敛速度快、稳定性高(和普通 K-means 相比)。本文用属性列的签名集 $P$ 和聚类数 $k$ 作为 K-means++的输入。输出为包含 $k$ 个候选匹配集的集合 $R$, $R=\{r_1,r_2,\cdots r_k\}$ ,其中 $r_i$ 为第 $i$ 个候选匹配属性集合。

随着 $k$ 值的变化,聚类结果可能会出现两种情况：a)表示不同语义的属性可能聚为一类;b)表示相同语义的属性可能存在于不同类中。针对问题 a)本文提出一种属性混合相似度计算方法来区分候选匹配集中语义不一致的属性。针对问题 b)本文运用启发式思想动态寻优 $k$ 值。

## 3 基于混合相似度的多模式匹配

为了更加准确地区分候选匹配集中的误配属性,提出一种新颖的基于相似度区分能力的混合相似度计算模型,进而根据该模型计算属性的混合相似度并构造属性映射图来描述属性之间的匹配关系。

### 3.1 混合相似度计算方法

#### 3.1.1 基于语法和语义的属性相似度计算

民航旅客服务数据的属性存在词干表达和复合词表达的形式,在基于 TF-IDF 计算不同模式属性的语法相似度之后需要对其进行拆分和词形还原的标准化处理,再计算其语义相似度。例如(TICKET_NO)->{ticket, number}。

a)基于 TF-IDF 的语法相似度计算[4]。首先将候选匹配集中的属性用 q-gram 方法分割成字母单元,再通过 TF-IDF 方法计算字母单元的权重 $w$,最后用一组权重向量 $\boldsymbol{v}=(w_1,w_2,\ldots,w_n)$来表示属性,属性 $sn_i$ 与 $sn_j$ 之间的语法相似度表示为:

$$\text{EdSim}(sn_i,sn_j)=\frac{v_i\cdot v_j}{\|v_i\|\times\|v_j\|} \tag{2}$$

b)基于 WordNet 的语义相似度计算[16]。在 WordNet 中影响两个概念词之间相似度的因素有两个,分别是两个概念词在 WordNet 中的距离和概念词在 WordNet 中包含的信息内容(IC)。其中影响 IC 值的因素有概念词的在 WordNet 中的深度和概念词在 WordNet 中的密度。IC 值与概念词的密度呈负相关与概念词的深度呈正相关。为准确表示属性之间的语义相似度,本文使用基于 IC 的相似度计算模型来衡量属性之间的语义相似度。计算模型定义如下。

$$\text{IC}(sn)=1-\frac{\log(\text{hypo}(sn)+1)}{\log(Node_{\max})}\times\frac{e^{\lambda\times\text{depth}(sn)}-e^{-\lambda\,\text{depth}(sn)}}{e^{\lambda\times\text{depth}(sn)}+e^{-\lambda\,\text{depth}(sn)}} \tag{3}$$

$$\text{L}(\text{IC})=\text{IC}(sn_i)+\text{IC}(sn_j)-2\times\text{IC}(sn_i,sn_j) \tag{4}$$

$$\text{L}(\text{path})=\frac{\log(\text{Dis}(sn_i,sn_j)+1)}{\log(2\times Depth_{\max}+1)} \tag{5}$$

$$\text{WtSim}(sn_i,sn_j)=e^{-(\alpha\times\text{L(IC)}+\beta\times\text{L(path)})} \tag{6}$$

其中:IC($sn$)表示属性 $sn$ 包含的信息内容,L(IC)表示 IC 语义距离 ,L(path) 表示两个概念词基于最短路径的语义距离,WtSim($sn_i$,$sn_j$)表示属性 $sn_i$ 和 $sn_j$ 之间的语义距离。hypo($sn$)表示属性 $sn$ 在 WordNet 中的下位词数量,$Node_{max}$ 表示 WordNet 中所有概念节点的数量,depth($sn$)表示 $sn$ 在 WordNet 中的深度,Dis($sn_i$,$sn_j$)表示属性 $sn_i$ 和 $sn_j$ 在 WordNet 中的最短距离。$\lambda, \alpha, \beta$ 为大于零的参数。

对于模式中非复合表达形式的属性按照式(6)计算其语义相似度,对于复合表达形式的属性经过还原处理后是由两个及以上的单词构成的词集时,先通过式(6)计算词集中单词的相似度,此时 WtSim($sn_i$,$sn_j$)为属性的词集相似度。

#### 3.1.2 混合相似度计算模型

由于单独使用语法或语义相似度无法准确表示属性之间的相似关系,因此结合两者提出一种新的混合相似度计算模型。

对于一个标注为匹配的属性对,若使用语法(语义)相似度计算方法得到的相似度值越接近 1,则可认为该种相似度的区分能力越强。对于一个标注为不匹配的属性对,若也使用该相似度计

算方法得到的相似度值越接近 0,则可认为该种相似度的区分能力越强。基于以上分析给出相似度区分能力定义。

定义 3　相似度区分能力。对于带标签属性对集合 $X^m$ 和 $X^u$,$X^m$ 中的属性对都被标注为匹配,$X^u$ 中的属性对都被标注为非匹配,$SIM_{X^m}$ 和 $SIM_{X^u}$ 分别表示基于语法(语义)相似度方法得出的相似度集合。则相似度区分能定义为:

$$\mathrm{dp}_{\mathrm{sim}}=\frac{\sum_{sim\in SIM_{X^m}} sim_i+\sum_{sim\in SIM_{X^u}} 1-sim_i}{|X^m|+|X^u|} \tag{7}$$

其中,$\mathrm{dp}_{\mathrm{sim}}$ 表示相似度区分能力,$\mathrm{sim}_i$ 表示属性对的相似度,$|X^m|$ 和$|X^u|$分别表示集合 $X^m$ 和 $X^u$ 中属性对的数量。

结合定义 3 可以得出基于相似度区分能力的混合相似度计算模型定义如下:

$$\mathrm{sim}(sn_i,sn_j)=\left(\begin{array}{l}\frac{\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})}{\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})+\mathrm{dp}_{\mathrm{sim}}(\mathrm{WtSim})}\times \mathrm{EdSim}(sn_i,sn_j)^p\\+\frac{\mathrm{dp}_{\mathrm{sim}}(\mathrm{WtSim})}{\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})+\mathrm{dp}_{\mathrm{sim}}(\mathrm{WtSim})}\times \mathrm{WtSim}(sn_i,sn_j)^p\end{array}\right)^{\frac{1}{p}} \tag{8}$$

其中:$\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})$和 $\mathrm{dp}_{\mathrm{sim}}(\mathrm{WtSim})$分别表示语法相似度和语义相似度的区分能力,$p$ 为参数,且 $p>0$。

### 3.2 基于混合相似度的多模式匹配算法

为了从候选匹配集中筛除误匹配项,获取最终的属性匹配关系,构建基于属性混合相似度的属性映射图,如算法 2 所示。

算法 2　构建属性映射图 $G(R,E)$

输入：匹配属性对集合 $X^m$ 和非匹配属性对集合 $X^u$,阈值 $\varepsilon$ 和 $\theta$,候选匹配集 $R=\{r_1,r_2,\ldots,r_k\}$。

输出：属性映射图 $G(R,E)$。

```
1 dp_sim(EdSim)'=dp_sim(WtSim)'=0.5
2 M= Ø, U= Ø
3 for x^m∈X^m,x^u ∈X^u
4      M=M.add(x^m),U=U.add(x^u)
5      if  |dp_sim(EdSim)-dp_sim(EdSim)'| < ε
6            return dp_sim(edsim)
7      else
8            dp_sim(EdSim)'= dp_sim(EdSim)
9      end if
10     if  |dp_sim(WtSim)-dp_sim(WtSim)'| < ε
11           RETURN dp_sim(WtSim)
12     else
13           dp_sim(WtSim)'= dp_sim(WtSim)
14     end if
15end for
16E= Ø
17for r∈R
18     for sn_i,sn_j∈r
19           if sim(sn_i,sn_j)>=θ
20                 E=E.add((sn_i,sn_j))
21           end if
22     end for
23end for
24return G(R,E)
```

具体步骤如下:

a)分别从标签为匹配和非匹配的集合 $X^m$ 和 $X^u$ 中每次挑选一个属性对分别加入集合 $M$ 和 $U$ 中,根据 $M$ 和 $U$ 用式(2)计算属性对的 TF-IDF 相似度 $\mathrm{EdSim}(sn_i,sn_j)$,进一步根据式(7)计算 TF-IDF 方法的区分能力 $\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})$,迭代这个过程直到 $\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})$的变化小于给定阈值 $\varepsilon$ 停止,并返回 $\mathrm{dp}_{\mathrm{sim}}(\mathrm{EdSim})$。同理可得 $\mathrm{dp}_{\mathrm{sim}}(\mathrm{WtSim})$值(3-15 行)。

b)遍历集合 $R$,先分别根据式(2)(6)计算 $r$ 中属性对$(sn_i,sn_j)$的 TF-IDF 相似度 $\mathrm{EdSim}(sn_i,sn_j)$和 WordNet 相似度 $\mathrm{WtSim}(sn_i,sn_j)$,进而根据式(8)计算其混合相似度 $\mathrm{sim}(sn_i,sn_j)$,若 $\mathrm{sim}(sn_i,sn_j)>=\theta$,则以属性 $sn_i$ 和 $sn_j$ 为顶点,将边$(sn_i,sn_j)$加入集合 $E$ 中,最后输出属性映射图 $G(R,E)$(17~24 行)算法结束。最后相互匹配的属性以边的形式连接,而误匹配的属性以孤立点的形式存在。属性之间的关系以图的形式直观的表示出来。

另外,通过分析 COMA 方法[2]发现影响多模式匹配时间复杂度的关键在于匹配过程中执行相似度计算的次数。参与匹配的模式数量以及模式中属性数量越多,需要执行相似度计算的属性对数量越多,COMA 方法时间复杂度就越高。设有 $n(n>=2)$个待匹配模式,每个模式平均包含了 $m(m>1)$个属性。在实际匹配过程中,模式中的属性存在冗余并不是一一对应,当选用 COMA 方法来处理多模式匹配时,匹配过程主要分为三步,首先从 $n$ 个模式中挑选两个模式,进而根据模式的属性类型选择对应的匹配器计算相似度,最后集成不同匹配器的结果。在该过程中只需关注执行相似度计算的属性对数量,将其设为 $Y_1$ 可得:

$$Y_1=\frac{1}{2}n*(n-1)*m^2 \tag{9}$$

本文的方法先对 $m*n$ 个属性聚类分析,获得 $k(k>1,k$ 为常数)个候选匹配集,进而根据候选匹配集中的属性计算相似度。设第 $i(i\le k)$个候选集中包含 $x_i$ 个属性,则有 $m*n=\sum_{i=1}^{k}x_i$ 。若将本文算法执行相似度计算的属性对数量设为 $Y_2$ 可得

$$Y_2=\sum_{i=1}^{k}\frac{1}{2}x_i*(x_i-1)=\frac{1}{2}\sum_{i=1}^{k}(x_i^2-x_i)=\frac{1}{2}\sum_{i=1}^{k}x_i^2-\frac{1}{2}\sum_{i=1}^{k}x_i \tag{10}$$

在这里 $x_i\ge 1$,当 $x_i=1$ 时表示候选匹配集中只有一个属性,不会进行相似度计算。实际情况下 $k\approx n$,设 $x_i=m$(实际情况下 $x_i$ 的取值在 $m$ 上下波动),则有 $Y_2=\frac{1}{2}n*m^2-\frac{1}{2}n*m$,由于 $Y_1>Y_2$,因此本文算法进行相似度计算的运算量更小。

同时根据以上分析可知,本文算法的时间复杂度为 $O(nm^2)$,而 COMA 方法的时间复杂度为 $O(n^2m^2)$,因此本文算法的时间复杂度更低。

综上所述,基于 SimHash 和混合相似度的多模式匹配方法

对应的处理流程如图 1 所示。

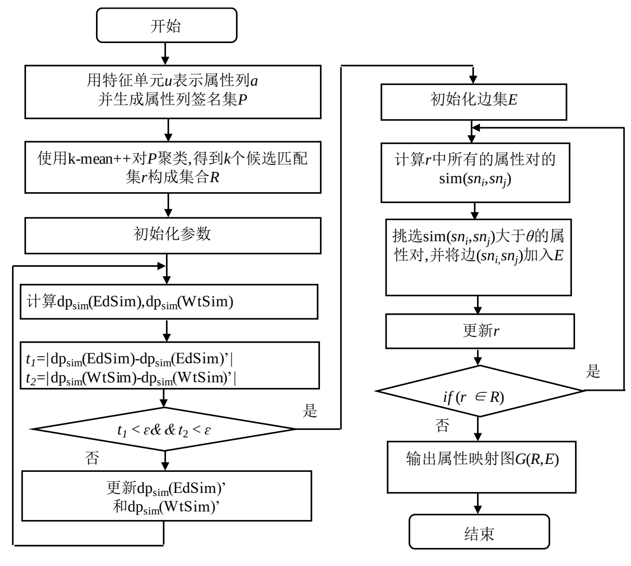



图 1 多模式匹配方法流程图

Fig.1 Flow diagram of multiple schema matching method

## 4 实验与评价

### 4.1 实验数据集

本文选用的实验数据来自于民航旅客服务系统(PSS)中的民航旅客订座数据(PNR)、电子客票数据(ET)、离港数据(CKI)和客座率数据(INV)四个数据源中的部分数据属性和大量数据实例。这些数据源的模式不同,而且各个系统又包含若干个功能模块,不同的功能模块为了提高查询速度设计了一些冗余的属性,这些属性虽名称不同,但却表示相同的语义。实验过程中将数据分成四组,代表四种来源的异构数据集。

表 3 属性及实例数量

Table3 Number of attributes and instances

| 异构数据源 | 属性数量 | 每个属性的实例数量 | 匹配的属性数量 |
|---|---|---|---|
| PNR | 16 | 30000 | 14 |
| ET | 11 | 30000 | 10 |
| CKI | 14 | 30000 | 14 |
| INV | 11 | 30000 | 10 |

### 4.2 评价指标

实验结果采用模式匹配领域中的查准率(Precision)、查全率(recall)和全面性(overall)三个指标进行评价。若 $T$ 为模式匹配算法返回的正确匹配结果数量;$P$ 为算法返回的所有匹配结果数量;$F$ 为算法返回的错误的匹配结果数量;$R$ 为实际所有正确的匹配结果数量。

查准率 $$\text{Precision} = \frac{T}{P}$$

查全率 $$\text{Recall} = \frac{T}{R}$$

全面性 $$\text{Overall} = \frac{T-F}{R} = \text{Recall} \times \left(2 - \frac{1}{\text{Precision}}\right)$$

### 4.3 实验结果及分析

实验过程主要从聚类 $k$ 值、相似度区分能力、综合相似度阈值 $\theta$ 、数据实例的数量这四个角度考虑不同因素对匹配结果的影响,并设计了多组对比实验,验证本文方法的可行性。

#### 4.3.1 k 值对聚类结果影响

聚类数 $k$ 值的选择对实验结果有一定的影响。若 $k$ 值选择过小,类与类之间就不易区分,影响最终匹配的查准率。若 $k$ 值选择过大,本属于同一类的属性会被分开影响最终的查全率。实验中数据实例数量为 500,聚类算法使用 kmeans++,实验结果如图 2 所示。

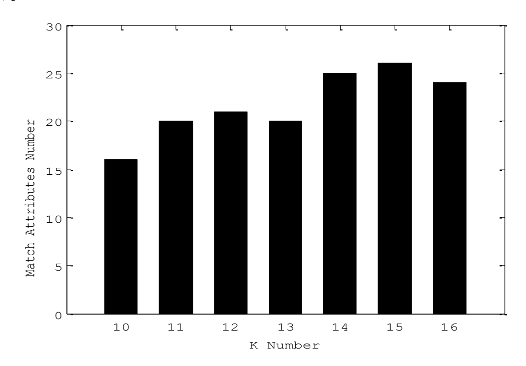



图 2 $k$ 取不同值时的聚类结果

Fig.2 Cluster results with different k values

根据图 2 可得,横坐标表示聚类数 $k$ 的取值,纵坐标表示 $k$ 取不同值时所有类中完全匹配的属性的数量之和。可以看出当 $k$=15 时完全匹配的属性数量达到峰值,聚类效果最好。而由表 3 可知单个模式的属性最大数量为 16,此时 $k$ 值与之相近。说明在实际匹配过程中 $k$ 取值应当与单个模式中属性的最大数量相近。

#### 4.3.2 探究不同相似度的区分能力

每次从待匹配属性对中随机挑选适量的属性对,构成匹配集合和不匹配集合。其中匹配集合与不匹配集合中的属性对数量相等。根据文献[16],WordNet 相似度的参数设置为 $\lambda$=0.4,$\alpha$=0.2,$\beta$=0.1,实验结果如图 3 所示。

根据图 3 可得横坐标表示匹配属性对集合与不匹配属性对集合中元素的数量之和,纵坐标表示相似度区分能力。从图中可知 WordNet 的区分能力值稳定在 0.69 左右,而 TF-IDF 的区分能力值呈上升到平稳的趋势,最后稳定在 0.61 左右。由于 TF-IDF 方法对同义词构成的属性不敏感,因此与 WordNet 方法相比,TF-IDF 方法的区分能力较弱。

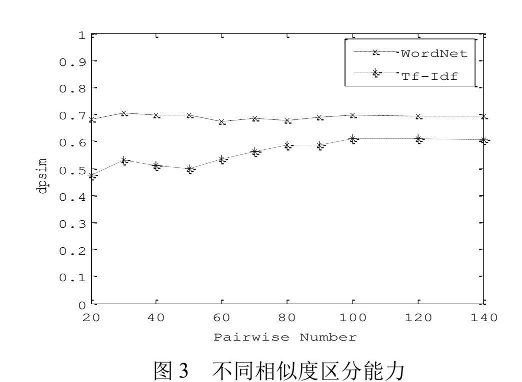



图 3 不同相似度区分能力

Fig.3 Distinguish ability of different similarity

4.3.3 阈值 θ 的选择对匹配结果的影响

构建属性映射图时,选择合适的阈值有助于提高最终匹配结果的准确性。实验中数据实例数量为 500,阈值区间为[0.3,0.6], $\theta$ 以间隔 0.05 递增,实验结果如图 4 所示。

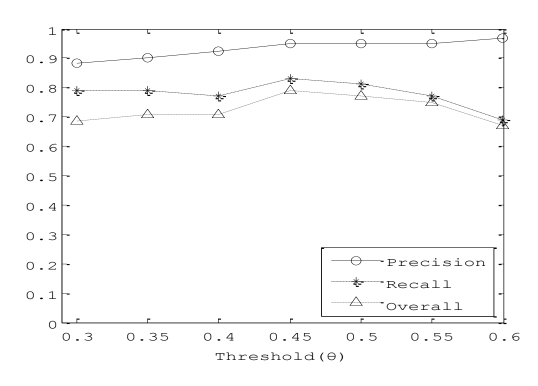



图 4 不同阈值下的匹配结果

Fig.4 Matching results with different thresholds($\theta$)

根据图 4 可得查准率总体处于一个上升的趋势。原因在于随着 $\theta$ 的增大,候选匹配集中的误匹配项被逐渐排除,所以查准率增大。查全率和全面性的总体趋势是先上升,后下降的趋势。当 $\theta$ 在 0.3~0.4 时,召回率在 $\theta$=0.4 时出现了下降,因为存在属性的相似度分布在不同区间的情况。而当 $\theta$ 取值为 0.45 时,查全率取值 0.832,达到峰值。当 $\theta$ 大于 0.45 时,更多正确匹配的属性经过 $\theta$ 筛选后被排除变成了孤立点,所以查全率开始下降。结合全面性指标走势可以得出当 $\theta$ 取值 0.45 时,通过属性映射图获得的匹配结果最优。

4.3.4 实例数量对匹配结果的影响

选用不同量级的数据实例进行对比实验,分析实例数量对匹配结果的影响。实验中 $k$ 取值 15,阈值 $\theta$ 取 0.45,实验结果如图 5 所示。

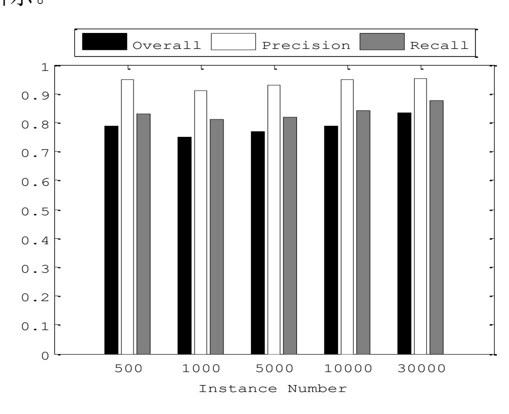



图 5 实例数量不同时的匹配结果

Fig.5 Matching results with different instances number

根据图 5 中可得三种指标总体上呈现一种微弱的上升趋势。当数据实例数量为 500 时的全面性指标为 0.793,当数据实例的数量为 1000 时的全面性指标值较小为 0.751。这主要是由于数据实例数量较少时实例分布不均导致的。而当数据实例的数量为 30000 时,全面性指标值达到 0.830。因此,参与匹配的数据实例数量越多,匹配结果的准确性越高。

4.3.5 不同方法对比实验分析

本文提出的基于 SimHash 和混合相似度的多模式匹配方法简称 B_SHM 方法。对比方法一来自文献[4]中的方法。该方法用 TF-IDF 算法提取属性的特征并聚类分析,从而获取属性的匹配关系。在这里简称其为基于模式信息的匹配方法(B_ATT)。对比方法二来自文献[10]中的方法。该方法用 q-gram 方法提取数据实例特征,并构造互信息向量,进而用聚类算法来求属性之间的匹配关系。在这里简称其为基于数据实例的匹配方法(B_INS)。

实验 1 本文提出的 B_SHM 方法的实验参数为 $k$ 取 15,阈值 $\theta$ 取值 0.45,数据实例数量为 30000,实验结果如表 4 所示。

表 4 三种不同方法对比实验结果

Table4 Comparison experimental results with the three different methods

| 实验方法 | Precision | Recall | Overall |
|---|---|---|---|
| B_SHM | 0。951 | 0。875 | 0。830 |
| B_ATT | 0。758 | 0。791 | 0。538 |
| B_INS | 0。813 | 0。875 | 0。674 |

实验 2 根据实验数据集,分别取 2 个模式,3 个模式和 4 个模式进行分组实验,测试不同方法的运行时间,实验结果如表 5 所示。

表 5 三种不同方法运行时间对比实验结果

Table5 Comparison experimental results of running time with three different methods

| 实验方法 | 2 个模式 | 3 个模式 | 4 个模式 |
|---|---|---|---|
| B_SHM | 0．303s | 0．412s | 0．547s |
| B_ATT | 0．135s | 0．216s | 0．324s |
| B_INS | 0．576s | 1．296s | 2．304s |

根据表 4 可得本文提出的 B_SHM 方法的查准率最高。B_INS 方法的查准率高于 B_ATT 方法。同时对比查全率指标可以得出 B_ATT 方法的查全率相对最低,而 B_SHM 方法和 B_INS 方法的查全率持平。最后结合全面性指标分析可知本文提出的 B_SHM 方法与 B_ATT 方法相比提高了 0.292,与 B_INS 方法相比提高了 0.156。这是因为 B_ATT 方法的匹配过程进使用属性的语法相似度来进行匹配,而不同模式的属性更可能发生"同名异义"或者"同义异名"的情况,所以匹配结果的查准率相对较低。B_INS 方法仅使用数据实例来辅助匹配,但存在属性语义不同,数据实例特征相似的情况,因此查准率也会受限,而 B_SHM 方法正好弥补了 B_ATT 方法和 B_INS 方法的不足。同时,根据表 5 可得当模式数量相同时,B_INS 方法的耗时最长,B_SHM 次之,B_ATT 方法耗时最短。原因在于 B_ATT 方法和 B_SHM 方法相比,B_SHM 方法在聚类处理后又进行了必要的相似度计算来排除误配属性,因此耗时较长。B_SHM 方法和 B_INS 方法相比,B_SHM 方法将高维的属性列特征转换成 128 位的指纹,因此聚类耗时远低于 B_INS 方法。通过综合全面性指标和时间性能可知,本文提出的方法在处理民航旅客服务数据的模式匹配过程中具有相对较好性能。






## 5 结束语

通过对多源异构民航旅客服务数据的分析,并针对现有模式匹配方法存在的效率低、精确度不足的问题,本文提出了基于 SimHash 和混合相似度的多模式匹配方法。首先,本文使用属性列的签名集进行聚类得到属性的候选匹配集,达到对属性初步筛选的目的。在很大程度上避免了同名异义"或者"同义异名"的问题。其次,本文提出了一种更准确的混合相似度计算模型,平衡了单一相似度计算带来的误差。最后,根据混合相似度确定属性映射关系,有效区分了因为数据实例特征相似导致误配的属性。同时该方法也避免了获取完整数据模式信息的繁琐过程。通过实验和分析表明本文提出的方法在多源异构数据集成领域可以有效解决多模式匹配效率低下和精确性不足的问题,在民航旅客服务数据集成方面具有重要的应用价值。

## 参考文献:


[1] 郑文怡, 鞠时光. 模式匹配方法研究 [J]. 计算机应用研究, 2006, 23 (2): 60-63. (Zheng Wenyi, Ju Shiguang. Research on schema matching approaches [J]. Application Research of Computers, 2006, 23 (2): 60-63. )

[2] Do Honghai, Rahm E. COMA: a system for flexible combination of schema matching approach [C]// Proc of the 28th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2002: 610-621.

[3] Li, WenSyan, Clifton, *et al.* SEMINT: a tool for identifying attribute correspondences in heterogeneous databases using neural networks [J]. Data & Knowledge Engineering, 2000, 33 (1): 49-84.

[4] Ding G, Sun T, Xu Y. Multi-schema matching based on clustering techniques [C]// Proc of the 10th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE Press, 2013: 778-782.

[5] Melnik S, Garcia Molina H, Rahm E. Similarity flooding: a versatile graph matching algorithm and its application to Schema matching [C]// Proc of the 18th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2002: 117-128.

[6] Madhavan J, Bernstein P A, Rahm E. Generic schema matching with cupid [C]// Proc of the 27th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers Inc. 2001: 49-58.

[7] 杜小坤, 李国徽, 王江晴, 等. 基于信息元的模式匹配方法 [J]. 软件学报, 2015, 26 (10): 2596-2613. (Du Xiaokun, Li Guohui, Wang Jiangqing, *et al.* Schema matching method based on information unit [J]. Journal of Software, 2015, 26 (10): 2596 (2613. )

[8] Bilke A, Naumann F. Schema matching using duplicates [C]// Proc of the 21th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2005: 69-80.

[9] Dhamankar R, Lee Y, Doan A, *et al.* IMAP: discovering complex semantic matches between database schemas [C]// Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2004: 383-394.

[10] Ahmadi B, Hadjieleftheriou M, Seidl T, *et al.* Type-based categorization of relational attributes [C]// Proc of the 12th International Conference on Extending Database Technology: Advances in Database Technology. New York: ACM Press, 2009: 84-95.

[11] Mehdi O, Ibrahim H, Affendey L. An approach for instance based schema matching with google similarity and regular expression [J]. International Arab Journal of Information Technology, 2017, 14 (5): 755-763.

[12] Gu B, Li Z, Zhang X, *et al.* The interaction between schema matching and record matching in data integration [J]. IEEE Trans on Knowledge & Data Engineering, 2016, 29 (1): 186-199.

[13] 石浩宏, 杨卫东. 基于全局本体的多数据源模式匹配方法的研究 [J]. 小型微型计算机系统, 2016, 37 (6): 1148-1152. (Shi Haohong, Yang Weidong. Research on method of multiple data sources schema matching based on global ontology [J]. Journal of Chinese Computer Systems, 2016, 37 (6): 1148-1152. )

[14] Manku G S, Jain A, Sarma A D. Detecting near-duplicates for web crawling [C]// Proc of the 16th International Conference on World Wide Web. New York: ACM Press, 2007: 141-150.

[15] Arthur D, Vassilvitskii S. K-means+: the advantages of careful seeding [C]// Proc of the 18th ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[16] 张思琪, 邢薇薇, 蔡圆媛. 一种基于 WordNet 的混合式语义相似度计算方法 [J]. 计算机工程与科学, 2017, 39 (5): 971-977. (Zhang Siqi, Xing Weiwei, Cai Yuanyuan. A WordNet-based hybrid semantic similarity measurement [J]. Computer Engineering and Science, 2017, 39 (5): 971-977. )